# 结合话题相关性的热点话题情感倾向研究*

何 跃 肖 敏 张 月
(四川大学商学院 成都 610064)

**摘要**: 【目的】热点话题具有很大的影响力,针对热点话题及其情感对象的情感倾向进行相关研究。【方法】提出一个结合话题相关性的主客观分类模型, 帮助抽取与热点话题相关的主观微博; 利用基于机器学习改进的情感分类方法对抽取博文的情感极性进行分析; 通过召回率、准确率、F 值对情感分类效果进行详细评估。【结果】实证分析结果表明, 结合话题相关性有效提升了热点话题微博主客观分类和情感极性分类效果, 其中 F 值分别提升 7.4%和 2.2%。【局限】待需深入考虑数据的分布状态、情感分类粒度细化、情感对象的情感趋势变化等。【结论】考虑话题相关性, 提升微博情感分类的效果, 并通过抽取热点话题中关键情感对象的情感倾向, 为微博精准营销提供相关情报信息。

**关键词**: 热点话题 主客观分类 情感倾向分类 TF-IDF-SIM 机器学习
**分类号**: G350



## 1 引 言

随着网络技术的发展, 互联网早已成为信息传播的一个重要载体, 微博、博客等社交网络凭借其丰富服务内容和便捷操作特色迅速融入人们的生活。尤其是 Web2.0 的出现, 微博成为这个时代具有强大影响力的产品, 它实现了信息的快速传播与交流, 对整个社会的政治、文化、经济各个方面也产生了重大影响, 越来越多的企业、商家、名人通过微博扩大知名度、提升公众形象。其中热点话题具有极大的影响力, 不仅影响着虚拟网络社会中各种事件的形成与发展, 同时也影响着真实人类社会中人们对于事件的看法和判断, 甚至于影响着政府与司法机构对事件的判决。所以面对微博中海量的文本数据, 快而准地抓住热点及焦点, 提取并分析用户的观点和情感信息, 对企业和政府来说都是非常有意义的。本文针对热点话题及其情感对象的情感倾向进行研究, 为微博精准营销提供情感方面的相关情报信息。

## 2 相关研究

微博随着其影响力和用户数量的不断扩大, 相关研究逐步增加, 针对微博情感倾向分析的文章也日益增多, 现有文献所用到的方法可以分为两大类: 基于情感词典和基于机器学习的方法[1]。

(1) 基于情感词典的方法通常是利用词典中词语的情感极性和强度, 对给定文本进行加权从而得到整个文本的情感倾向, 目前常见的中文情感词典有 HowNet、NTUSD 情感词典、学生褒贬义词典和 Tsinghua 褒贬义词典等。基于情感词典的方法不需要训练数据且可以应用于很多领域, 但是它在微博情感分析中仍有以下缺陷:

①情感词典的获取成本及更新成本较大, 桂斌等[2]基于微博表情符号提出一种自动构建情感词典的方法, 但这种方法只能应用于对微博情感分析要求不是特别高的场合; Bravo-Marquez 等[3]提出以监督的方式从表情符号自动标注的 Twitter 和已有词典来扩大词典, 其中使用点互信息和随机梯度下降法建立词语和情感之间的联系, 实验结果表明提升了 SentiWordNet(英文情感分析词典)的性能, 利用表情符号自动标注的数据虽然能降低成本, 但也会降低方法的有效性。

②情感词典中词语的覆盖率较低, 使得微博中一些新兴词汇、错误拼写词汇、缩写词汇、非正式词汇等难以被覆盖, 宁慧等[4]将多个词典通过合并去重等方式构建一个新的

通讯作者: 肖敏, ORCID: 0000-0002-6508-3551, E-mail: 2279332915@qq.com。
*本文系四川大学中央高校基本科研业务费项目"基于中文微博的负面情绪预警研究"(项目编号: skqy201406)的研究成果之一。

词典，但微博中词语日新月异，随时会出现新的具有强烈情感的词汇; Zhou 等[5]通过加入与具体领域相关的意见词汇扩大情感词典，在 56 个话题上的实验结果表明可以提升基于词典的分类器的准确率。

③词语固定的情感极性和强度使得该方法是领域无关的，而情感表达时通常涉及到具体的对象或领域，在不同的语境中具有不同的情感强度，情感词典的领域无关性对情感分类的影响特别大，近年来越来越多的学者关注于解决这个问题，如 Saif 等[6-8]进行了一系列研究，提出 SentiCircles 方法，通过词语的共现模式，动态更新应用在具体数据集时情感词典中词语的情感分数值，这个方法充分考虑到了词语出现的语境，能提高在特定领域的情感分类结果，但是在交叉领域时没有明显改善; 此外还提出从 DBpedia 抽取出语义关系来提高词典的适应性[9]，结果表明有效提高了情感分类的准确率和 F 值; Zhao 等[10]结合语义和先验情感也取得了较好的性能。

(2) 基于机器学习的方法要求训练数据用于情感分类学习，训练数据通常是人工标注微博的情感倾向(积极、消极和中立等)，目前常用的方法有支持向量机、朴素贝叶斯、神经网络、最大熵等[11-12]，其中支持向量机在许多文献中被证明分类结果较好。基于机器学习的方法依赖于训练数据，因此在分析特定微博话题的情感时往往更具有优势，但是仍有以下缺陷:

①训练数据获取成本较高，利用文本中的表情符号自动标注数据是一种常见的方法[3-4]，但是标注结果仍然有待提高。

②训练数据的选择对情感分类的结果影响特别大，Palguna 等[13]分析了 Twitter 的抽样算法并提出新的统计指标来量化样本的代表性; Song 等[14]认为微博中的情感表达反映了用户的个性，训练数据中用户的代表性也是一个值得关注的问题; 另外训练数据集的大小对结果也有影响[15]。

③训练数据的平衡性对于分类器也会产生影响，目前相关研究多是通过对抽样方法进行改进来解决这个问题[16-17]。在基于机器学习的方法用于微博话题的情感分析时，由于基于监督学习的方法自身是领域相关的，很多之前的研究往往忽略了微博内容与话题的相关性，从而使得部分噪声数据降低了训练结果的有效性，这也是本文的主要切入点。

这两类方法各有千秋，目前亦有许多研究结合两者的优势来研究微博的情感倾向[1,18]。笔者认为，在研究特定热点话题的情感倾向时基于机器学习的方法更具有优势，所以本文对微博热点话题的情感倾向进行改进研究，充分考虑噪声数据的影响，即博文内容与话题的相关性强度。本文微博话题情感倾向研究主要有三大任务: 文本预处理、情感信息抽取和情感分类。首先利用文本预处理技术对热点话题微博进行情感信息(如特征词、情感对象)的抽取，然后提出结合话题相关性的主客观分类模型来帮助抽取与热点话题相关且主观的微博文本集合，并利用改进的主观微博情感分类方法对微博的情感倾向进行分析，最后通过召回率、准确率、F 值对分类效果进行详细评估。在实证分析中，对热点话题#冯小刚炮轰影评人#的相关情感对象的情感倾向进行研究分析，并针对结果提出一些微博精准营销的建议。

## 3 研究设计

### 3.1 数据获取及文本预处理

为了研究微博热点话题的情感倾向，首先通过网页爬虫软件获取相关数据，并对数据进行适当的预处理。其中预处理程序包括: 提取微博中表情符号; 清洗无意义的微博文本，包括纯粹的转发微博、图片、视频、网址、表情、URL 地址等; 分词及词性标注; 过滤停用词等。

### 3.2 情感信息抽取

(1) 特征词抽取

抽取特征词主要是从文本中抽取出能代表文本内容且对其分类起决定性作用的词，并计算其特征权重。目前常见的特征抽取方法有文档频率 DF、信息增益 IG、互信息 MI 和卡方检验 CHI 等。不同分类下使用到的特征抽取方法有所差别，本文涉及到三种分类，包括话题相关性分类、主客观文本分类和情感极性分类。其中话题相关性分类特征包括词、词性及其与话题的相似度值，采用基于 TF-IDF 改进的 TF-IDF-SIM 算法确定各个特征词的权重。张想[19]在主客观分类常用的五维非文本特征上，加入了三维新的特征，利用 4 种常用分类器(SVM、ANN、NB 和 LR)对比五维特征和八维特征(如表 1 所示)的效果，证明八维特征效果更好，进一步探究表情符号特征对主客观分类的影响，发现表情符号特征有效提升了分类效果。吴青林等[20]构建的一个较为完整情感极性词典中包括基础情感词典、极性副词词典、表情词词典、微博新词词典和领域词典，结合文献，笔者认为如表 2 所示的 6 类特征项是微博中最常出现的情感特征，并采用互信息 MI 抽取，其中表情符号的存储形式为“[文字]”，将其提取出来后按照其在 HowNet 情感词典中的情感进行

分析，网络用语是从网站上人工收集；在情感极性进行分类时，在表2的基础上，将表情符号、情感词及网络用语分为正负两类，并加入转折词，需要注意网络用语的情感极性采用人工标注的方式，最终得到如表3所示的10类特征项，并采用互信息MI抽取。

表1 八维主客观分类特征[19]

| 对比项 | 特征 | 取值 |
| --- | --- | --- |
| 常用的五维分类特征 | 是否含有情感词 | 0, 1 |
| | 是否含有感叹号 | 0, 1 |
| | 是否含有问号 | 0, 1 |
| | 是否含有主张词 | 0, 1 |
| | 是否含有程度副词 | 0, 1 |
| 张想[19]加入的三维新特征 | 是否含有代词或名词 | 0, 1 |
| | 微博句子数目 | Real |
| | 微博所含词的个数 | Real |

表2 主客观文本分类特征

| 特征类型 | 特征内容 | 描述 | 特征取值 |
| --- | --- | --- | --- |
| 表情符号 | 情感表情符号个数 | 新浪微博默认表情类 | Real |
| 情感词 | 情感词出现个数 | HowNet情感分析用词语集 | Real |
| 网络用语 | 网络用语词出现个数 | 人工收集的网络用语词典，含褒义词和贬义词 | Real |
| 否定词 | 是否出现否定词 | 是否情感词前面存在否定词(否定词23个，来源是HowNet词典) | 0, 1 |
| 程度副词 | 是否含有程度副词 | HowNet词典中的程度词词典 | 0, 1 |
| 语气词 | 是否含有语气词 | “呀”、“啦”、“呢”、“吧”、“啊”等25个语气词 | 0, 1 |

(2) 情感对象抽取与合并

情感对象，即评价对象，是指在主观句中情感词或短语修饰的词，可以是个人、组织、事件和产品等对象。情感对象的抽取及其情感倾向判断有助于微博精准营销，目前微博情感对象抽取方法有基于规则的方法、基于句法分析的方法和序列标注模型的方法等，其中基于规则的方法比较简单且效率较高，故采用此方法进行评价对象抽取。在评价对象中，名词、名词短语及话题标签(Hashtag)占主要部分，如果有具体名词或名词短语，则将其作为评价对象，否则将话题标签作为评价对象。然而抽取出的评价对象往往存在大量的相似词汇，如“冯小刚”和“冯导”表示相同的意思，

表3 情感极性分类特征

| 特征类型 | 特征内容 | 描述 | 特征取值 |
| --- | --- | --- | --- |
| 正面表情符号 | 正面表情符号个数 | 新浪微博默认表情类 | Real |
| 负面表情符号 | 负面表情符号个数 | 新浪微博默认表情类 | Real |
| 正面情感词 | 正面情感词个数 | HowNet中的正面情感词 | Real |
| 负面情感词 | 负面情感词个数 | HowNet中的负面情感词 | Real |
| 正面网络用语 | 正面网络用语词个数 | 褒义的网络用语词典 | Real |
| 负面网络用语 | 负面网络用语词个数 | 贬义的网络用语词典 | Real |
| 否定词 | 是否出现否定词 | 是否情感词前面存在否定词(情感词前3个词之内) | 0, 1 |
| 程度副词 | 是否含有程度副词 | HowNet词典中的程度词词典 | 0, 1 |
| 语气词 | 是否含有语气词 | “呀”、“啦”、“呢”、“吧”、“啊”等25个 | 0, 1 |
| 转折词 | 是否含有转折词 | “但是”、“可是”、“然而”等7个常用词 | 0, 1 |

故本文采用基于相似度计算和词语覆盖率的K-means聚类算法对评价对象进行合并。

### 3.3 结合话题相关性的主客观分类模型

在研究话题情感倾向之前，需要抽取与话题相关且主观的博文，因为主观文本内容基于断言或评论且带有个人情感和意向的抒发，而客观文本内容基于事实描述且不带有个人的好恶和偏见。本文提出一个结合话题相关性的主客观分类模型，将问题分解为两个并行子问题，即是否相关和是否主观，然后利用Logistic回归进行归并，从而得到与热点话题相关的主观博文。结合话题相关性的主客观分类模型如图1所示。

从图1可以看出，基于话题相关性的主客观分类模型包含两个子模型，即话题相关性分类子模型和主客观分类子模型。两个子模型的主要流程均包括特征项抽取、特征矩阵建立、样本序列建立和模型学习4个阶段，其中样本序列建立阶段均使用到人工标注方法，模型学习阶段均使用了SVM算法。

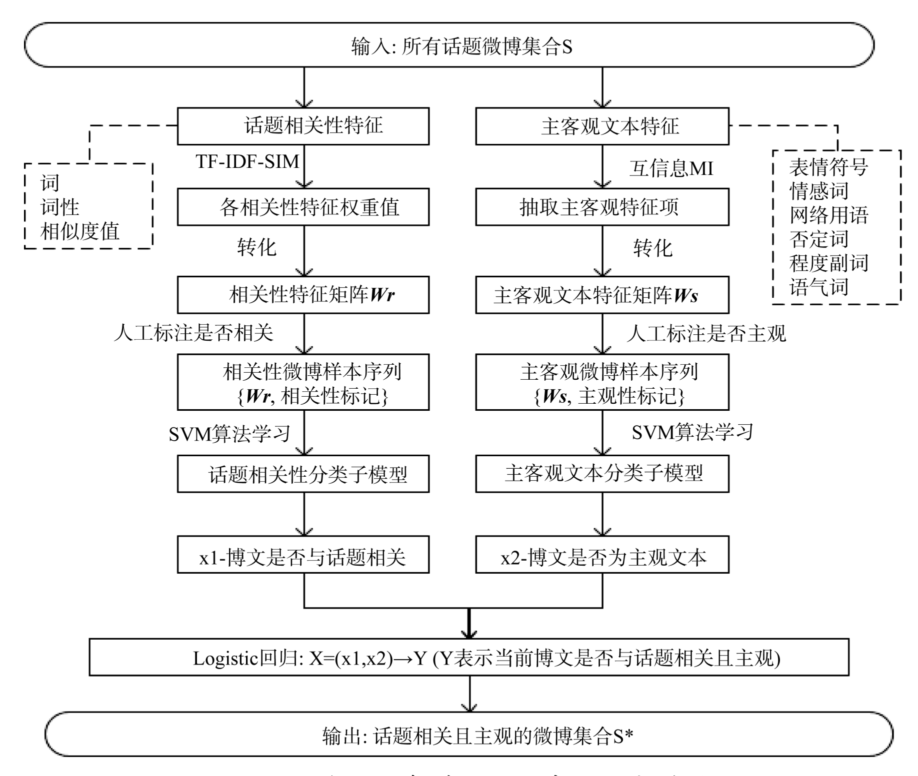



图 1　结合话题相关性的主客观分类模型

特征矩阵建立阶段, 话题相关性子模型使用的是 TF-IDF-SIM 法, 主客观分类子模型中使用的是互信息 MI 法。

(1) TF-IDF-SIM 法, 是基于 TF-IDF 改进的算法, 综合考虑一个术语对特定话题语料库的重要程度及与特定话题的相关程度, 从而为这个术语赋予一个综合权重, 其中 TF(Term Frequency)表示词频, IDF(Inverse Document Frequency)表示反文档频率, SIM 表示术语与话题词的最大相似度值。在文本特征表示时, 每条博文 $D_j$ 都可以用博文中词的特征来表示, 这些词的特征及其权重就构成"空间"中的向量:

$$(W_{1,j}, W_{2,j}, \cdots, W_{n,j}) \quad (1)$$

其中, $W_{i,j}$ 为词条 $i$ 在博文 $D_j$ 中的权重, 表示为:

$$W_{i,j} = TF_{i,j} \times IDF_i \times SIM_i \quad (2)$$

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (3)$$

其中, $TF_{i,j}$ 表示词条 $i$ 在博文 $D_j$ 中的出现次数; $IDF_i$ 是反文档频率系数, $N$ 表示语料库中所有的博文条数, $n_i$ 表示语料库中出现过词条 $i$ 的博文条数; $SIM_i$ 为词条 $i$ 与当前话题词的相似度值, 相似度算法如下。

输入: 当前词 w 和当前热点话题词 hotTopic;

输出: 词 w 与当前热点话题词的最大相似度 sim(w, hotTopic);

①记 sim(w, hotTopic)=0;

②将 hotTopic 分词得到 hotTopicSet={H1, H2, …, Hn};

③对于每一个 H_i, 如果 w=H, 则 sim(w, hotTopic)=1, 转向步骤⑤; 否则转向步骤④;

④计算 sim(w, H_i)//利用基于同义词词林的词语相似度计算算法[21]得到

如果 sim(w, H_i)> sim(w, hotTopic), 则更新 sim(w, hotTopic)= sim(w, H_i);

⑤算法结束。

(2) 互信息 MI 法。将表 2 中的表情符号、情感词等 6 类情感特征抽取出来后, 分别计算它们与主观类和客观类文本的互信息(MI)。互信息(Mutual Information)是信息论里一种有用的信息度量, 它是指两个事件集合之间的相关性, 通过公式(4)计算, 其意义是由于事件 A 发生与事件 B 发生相关联而提供的信息量。

$$I(A,B) = \log_2 \frac{P(AB)}{P(A)P(B)} \quad (4)$$

其中, $P(AB)$表示事件 A 和事件 B 同时发生的概率, $P(A)$表示事件 A 发生的概率, $P(B)$表示事件 B 发生的概率。在处理分类问题提取特征的时候用互信息来

衡量某个特征项和特定类别的相关性，如果信息量越大，那么特征和这个类别的相关性越大，反之亦然。互信息法用于特征提取的基本流程是：假设特征项为 t，主观类是 $c_1$，客观类为 $c_2$，计算的结果为 $MI(t,c_1)$ 和 $MI(t,c_2)$，定义一个阈值 $\theta$，假如特征项满足公式(5)，则将该特征项抽取出来。

$$|MI(t,c_1)-MI(t,c_2)|>\theta \tag{5}$$

最后通过 Logistic 回归来组合两个子模型，从而构建一个结合话题相关性的主客观分类模型。从上述分析可看出，热点话题主客观分类问题的本质是寻找一个随机变量 $Y$ 与随机向量 $X=(x1,x2)$ 之间的函数关系，其中 $Y$ 代表当前言论是否为与话题相关且主观性的言论；$x1$ 为博文是否与话题相关的自变量、$x2$ 为博文是否为主观文本的自变量。因标记 $(x1,x2)$ 以及分类结果 $Y$ 都为离散型数据，选用 Logistic 回归分析解决上述的问题。

### 3.4 基于机器学习改进的主观微博情感分类

得到与话题相关的主观文本后，需要对主观文本进行情感倾向分类，并根据抽取出的评价对象判断每个评价对象的情感倾向。目前针对中文微博的情感分类方法可以分为两类：基于语义词典的情感计算法；基于机器学习的情感分类法。微博文本情感分析领域还没有一部通用且完整的情感词典，同时受语境迁移的影响，现有大多数情感词典在微博情感分析中都存在情感覆盖面不足、分类效果差的缺点[14-15]。故本文采用基于机器学习的方法，使用情感词、表情符号、网络用语等作为分类特征，通过分类算法训练一个分类器，将情感倾向分为正面倾向和负面倾向这两类。

本文改进了以往的主观微博情感分类特征。之前的许多研究提出将能表达情感的词汇(如名词、形容词、副词、动词等)作为特征项，但是未将非规范文本如表情符号、网络用语等考虑进去，而这些文本又是互联网时代人们表达情感的重要因素。采用如表 3 所示的情感极性分类特征，其中很多与主客观文本分类的特征相似，但把特征项：表情符号、情感词、网络用语词典细分为正面和负面两个方面，并且加入转折词。因为朴素贝叶斯算法对文本的适应性较强，对正面和负面倾向分类的整体效果稳定，所以利用朴素贝叶斯算法构建主观微博情感分类器。

## 4 实证分析

### 4.1 微博数据获取与预处理

使用火车采集器获取新浪微博热点话题#冯小刚炮轰影评人#中的数据总共91 361条，对文本进行预处理后剩下88 571条博文，将其随机拆分为训练集(68 889条)和测试集(19 682条)，训练集是测试集的3.5倍。然后使用人工标注法对微博文本进行标注，请三位专家分别从相关性上将博文标注为相关或无关，从情感极性上标注为客观、积极或消极三类。人工标注前给每位专家一个小册子，用于向专家解释相关分类的概念，方便专家参考，比如积极文本里通常会包含用户对事件的支持、看好的态度等。各专家相互同意度均超过80%，这表明通过了信度检验，最后根据频数最大法(即服从大多数人的意见)得到最终分类结果。分类标注结果如表4所示。

表 4　分类标记结果(单位:条)

| 情感极性 | 主观且相关 | | 其余 | | |
|---|---|---|---|---|---|
| | 正面 | 负面 | 主观且无关 | 客观且相关 | 客观且无关 |
| 合计 | 38 022 | 24 598 | 10 596 | 11 071 | 4 284 |
| | 62 620 | | 10 596 | 15 355 | |

### 4.2 热点话题的主客观分类

根据 2.3 节中的研究方法提取特征词及矩阵，在 WEKA 平台使用 SVM 分类器对话题相关性分类子模型和主客观文本分类子模型分别进行标注，再利用 Logistic 算法将两重标注统一在一个模型中，结果如表 5 所示。

表 5　主客观文本分类结果

| 对比项 | | | 数量/条 | 准确率(%) | 召回率(%) | F 值(%) |
|---|---|---|---|---|---|---|
| SVM 分类 | 主观文本 | 话题相关 | 53 356 | 82.5 | 89.3 | 85.8 |
| | | 话题无关 | 10 127 | 76.7 | 93.3 | 84.2 |
| | 客观文本 | 话题相关 | 15 365 | 68.5 | 73.8 | 71.1 |
| | | 话题无关 | 9 723 | 53.9 | 55.6 | 54.7 |
| Logistic 回归 | | 话题相关且主观文本 | 53 285 | 83.6 | 89.0 | 86.2 |

可以看出，在话题相关性分类时，SVM 分类器对主观文本比客观文本的分类效果更好，其准确率、召回率、F 值均比客观文本高，其“话题相关”类别的 F 值比客观文本高 14.7%，表明对于被判断为客观的文

本，分类器更容易将其判别为与话题无关，之后的研究可以进一步加强对客观文本的相关性识别的探索。利用 Logistic 回归模型得到的话题相关且主观的文本，准确率提高了 1.1%，F 值提高了 0.4%，说明 Logistic 回归模型在一定程度上提升了热点话题的主客观分类问题的效果，但是提升效果还不是很明显。

此外，研究是否加入话题相关性分类子模型对热点话题主客观文本分类的影响，结果如表 6 所示。可以发现，引用话题相关性分类子模型对热点话题的主客观文本分类效果更好，总体 F 值提高了 7.4%，这说明话题数据中与热点话题无关的言论文本影响了主客观文本分类的效果，通过结合话题相关性较大程度提高了主客观分类效果。

表 6　是否加入话题相关性分类子模型对热点话题主客观文本分类的影响

| 对比项 | 主观(%) | | | 客观(%) | | | 总体(%) |
|---|---|---|---|---|---|---|---|
| | 准确率 | 召回率 | F 值 | 准确率 | 召回率 | F 值 | F 值 |
| 未加话题相关性分类子模型 | 76.8 | 94.1 | 84.6 | 66.6 | 42.1 | 51.6 | 72.3 |
| 加入话题相关性分类子模型 | 88.2 | 92.3 | 90.2 | 81.5 | 53.8 | 66.8 | 79.7 |

## 4.3　主观微博的情感倾向分类

根据 3.4 节中基于机器学习改进的主观微博情感分类的方法，对抽取出的与话题相关且主观的微博的情感极性进行判断，并将改进前和改进后的结果进行比较，如表 7 所示。

表 7　话题情感倾向分类结果对比

| 对比项 | 情感倾向 | 数量(条) | 准确率(%) | 召回率(%) | F 值(%) |
|---|---|---|---|---|---|
| 改进前 | 正面倾向 | 34 479 | 80.5 | 87.6 | 83.9 |
| | 负面倾向 | 18 806 | 73.2 | 79.1 | 76.0 |
| 改进后 | 正面倾向 | 33 941 | 84.3 | 90.3 | 87.2 |
| | 负面倾向 | 19 344 | 79.8 | 77.6 | 78.7 |

从表 7 看出加入正负网络用语、正负表情符号等特征词对分类效果有所提升，正面倾向的 F 值提高了 3.3%，负面倾向的 F 值提高了 2.7%，另外在改进前和改进后，正面倾向的 F 值均比负面倾向的 F 值高，究其原因，可能是由于数据自身的不平衡性导致，数据集中正面倾向的数量较大程度大于负面倾向的数量。

此外还研究了是否加入热点话题的主客观分类模型对情感倾向分类的影响，结果如表 8 所示。

表 8　是否加入结合话题相关性的主客观分类模型对情感倾向分类的影响

| 对比项 | 正面(%) | | | 负面(%) | | | 总体 |
|---|---|---|---|---|---|---|---|
| | 准确率 | 召回率 | F 值 | 准确率 | 召回率 | F 值 | F 值 |
| 未加结合话题相关性的主客观分类模型 | 81.5 | 92 | 86.4 | 67.9 | 82.8 | 74.6 | 81.7 |
| 加入结合话题相关性的主客观分类模型 | 84.3 | 90.3 | 87.2 | 79.8 | 77.6 | 78.7 | 83.9 |

加入结合话题相关性的主客观文本分类模型对情感倾向分类的效果有所提升，总体 F 值提高了 2.2%，说明通过筛选出的相关且主观文本来进行情感分类，能够大大降低分类器的负担，从而提供更精确的分类效果。

## 4.4　情感对象抽取及情感倾向判断

根据 3.2 节的方法抽取与合并情感对象，并对评价对象讨论最多的前 5 名进行情感倾向判断，如表 9 所示。

表 9　情感对象及其情感倾向

| 对比项 | Hashtag | 冯小刚 | 私人订制 | 小故事 | 葛优 |
|---|---|---|---|---|---|
| 正面情感数量(条) | 13 526 | 7 158 | 2 330 | 417 | 532 |
| 负面情感数量(条) | 7 945 | 4 415 | 3 052 | 508 | 365 |

从表 9 可以看出，Hashtag(#冯小刚炮轰影评人#)的正面情感数量大约是负面的 1.7 倍，说明新浪微博用户对于这个事件的支持数量是远大于反对数量的。对评价对象“冯小刚”的正面情感数量大约是负面的 1.6 倍，而“小故事”和“葛优”是电影“私人订制”里的情节和演员，用户对这部电影和情节本身的负面情感较多。说明用户对电影的负面情绪并没有影响到大家对冯小刚的喜爱和这次事件的支持；从总体的情感倾向数量上来看，用户对话题的讨论还是更集中在对热点话题本身和冯小刚这两个对象上。

从微博营销的角度，电影制片商等企业可以得到用户态度、偏好反馈，并可以针对有负面评价的对象如私人订制等，在后期宣传上积极渲染电影的立意及电影背后的内涵，让观众了解更多有关电影正面的内容，从而进行舆论引导。政府部门也可以在出现大量

的负面信息之后，进行重点监督、排除敏感信息，从受关注最多的几个评价对象着手，制定舆论引导与控制的相关策略。

## 5 结　语

针对热点话题及其情感对象的情感倾向进行相关研究，提出一个结合话题相关性的主客观分类模型，帮助抽取与热点话题相关的主观微博；利用改进的情感分类方法对抽取博文的情感倾向进行分析；通过召回率、准确率、F 值对情感分类效果进行详细评估。实证结果表明：基于话题相关性的主客观分类模型，有助于热点话题的主客观分类，使得微博情感分类效果更好；通过抽取热点话题中关键情感对象的情感倾向，能为微博精准营销提供相关情报信息。

主要有以下创新：抽取了话题中的评价对象并进行情感倾向判断，比单纯的情感分类理论和技术探讨更具有价值；提出结合话题相关性的主客观分类模型，使得主客观文本分类效果有所提升，从而也提升了情感极性的分类效果，究其原因，可能是考虑到相关性之后降低了噪声数据产生的影响；提出改进的情感分类方法，在文本处理时考虑了非规范性文本如表情符号等，结果证明了该方法和模型的有效性和实用性。

仍存在一些值得改进之处：样本分布的不均匀可能会影响分类的效果，但由于本文研究的是一个话题内所有的微博数据，所以没有对样本分布进行调整；对情感倾向的分类仅分为正负两面，实际应用中更需要将情感类型的粒度细化；缺乏对情感对象进行情感趋势分析。在未来的研究中，可以针对这些问题进行深入研究。

## 参考文献：

[1] 陈国兰. 基于情感词典与语义规则的微博情感分析[J]. 情报探索, 2016(2): 1-6. (Chen Guolan. Microbiog Sentiment Analysis Basing on Emotion Dictionary and Semantic Rule[J]. Information Research, 2016(2): 1-6.)

[2] 桂斌, 杨小平, 张中夏, 等. 基于微博表情符号的情感词典构建研究[J]. 北京理工大学学报, 2014, 34(5): 537-541. (Gui Bin, Yang Xiaoping, Zhang Zhongxia, et al. Research on Building Lexicon for Sentiment Analysis Based on the Chinese Microblogging Smiley[J]. Transactions of Beijing Institute of Technology, 2014, 34(5): 537-541.)

[3] Bravo-Marquez F, Frank E, Pfahringer B. Building a Twitter Opinion Lexicon from Automatically-annotated Tweets[J]. Knowledge-Based Systems, 2016, 108(SI). DOI: 10.1016/j.knosys.2016.05.018.

[4] 宁慧, 杨松, 赵勇, 等. 基于语义特征的微博情感分析研究[J]. 应用科技, 2016, 43(3): 70-74. (Ning Hui, Yang Song, Zhao Yong, et al. Study of Microblog Sentiment Analysis Based on Semantic Feature[J]. Applied Science and Technology, 2016, 43(3): 70-74.)

[5] Zhou Z, Zhang X, Sanderson M. Sentiment Analysis on Twitter Through Topic-Based Lexicon Expansion[A]// Databases Theory and Applications[M]. Springer International Publishing, 2014:98-109.

[6] Saif H, Fernandez M, He Y, et al. SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter[A]// The Semantic Web: Trends and Challenges[M]. Springer, Cham, 2014: 83-98.

[7] Saif H, He Y, Fernandez M, et al. Adapting Sentiment Lexicons Using Contextual Semantics for Sentiment Analysis of Twitter[A]// The Semantic Web: ESWC 2014 Satellite Events[M]. Springer, Cham, 2014: 54-63.

[8] Saif H, He Y, Fernandez M, et al. Contextual Semantics for Sentiment Analysis of Twitter[J]. Information Processing & Management, 2015, 52(1): 5-19.

[9] Saif H, Fernandez M, Kastler L, et al. A Linked Open Data Approach for Sentiment Lexicon Adaptation [C]// Proceedings of the 15th International Semantic Web Conference. 2016.

[10] Zhao J, Cao X. Combining Semantic and Prior Polarity for Boosting Twitter Sentiment Analysis[C]//Proceedings of the 2015 IEEE International Conference on Smart City/Socialcom/Sustaincom. IEEE, 2015:832-837.

[11] Le B, Nguyen H. Twitter Sentiment Analysis Using Machine Learning Techniques[A]// Advanced Computational Methods for Knowledge Engineering [M]. Springer International Publishing, 2015: 279-289.

[12] Qasem M, Thulasiram R, Thulasiram P. Twitter Sentiment Classification Using Machine Learning Techniques for Stock Markets[C]//Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics. IEEE, 2015.

[13] Palguna D, Joshi V, Chakaravarthy V, et al. Analysis of Sampling Algorithms for Twitter[C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. AAAI Press, 2015.

[14] Song K, Feng S, Gao W, et al. Personalized Sentiment Classification Based on Latent Individuality of Microblog Users[C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. AAAI Press, 2015.

[15] Abdelwahab O, Bahgat M, Lowrance C J, et al. Effect of Training Set Size on SVM and Naive Bayes for Twitter

Sentiment Analysis[C]// Proceedings of the IEEE International Symposium on Signal Processing and Information Technology. 2015: 46-51.

[16] Saif H, He Y, Alani H, et al. On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter[C]// Proceedings of the 9th International Conference on Language Resources and Evaluation. 2014.

[17] Ah-Pine J, Morales E P S. A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis[C]// Proceedings of the Workshop on Interactions Between Data Mining and Natural Language Processing. 2016.

[18] Sabariah M K, Effendy V. Sentiment Analysis on Twitter Using the Combination of Lexicon-based and Support Vector Machine for Assessing the Performance of a Television Program[C]//Proceedings of the International Conference on Information and Communication Technology. 2015.

[19] 张想. 面向热点话题型微博的情感分析研究[D]. 哈尔滨: 哈尔滨工业大学, 2013. (Zhang Xiang. Research on Sentiment Analysis for Hot Topic Microblog[D]. Harbin: Harbin Institute of Technology, 2013.)

[20] 吴青林, 王焱. 中文微博情感特征选择方法研究[J]. 内蒙古师大学报: 自然汉文版, 2016, 45(1): 84-88. (Wu Qinglin, Wang Yan. Research on the Emotional Feature Selection Method in the Chinese Microblog[J]. Journal of Inner Mongolia Normal University: Natural Science Edition, 2016, 45(1): 84-88.)

[21] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602-608. (Tian Jiule, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System[J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.)

## 作者贡献声明:

何跃: 提出研究思路, 设计研究方案;
肖敏, 张月: 进行实验, 采集、清洗和分析数据;
何跃, 肖敏, 张月: 论文起草;
何跃, 肖敏: 论文最终版本修订。

## 利益冲突声明:

所有作者声明不存在利益冲突关系。

## 支撑数据:

支撑数据由作者自存储, E-mail: 2279332915@qq.com。

[1] 何跃, 肖敏, 张月. Data.xlsx. 原始数据及其人工分类标注结果表.

[2] 何跃, 肖敏, 张月. Data.xlsx. 主客观分类结果表.

[3] 何跃, 肖敏, 张月. Data.xlsx. 是否加入话题相关性分类子模型时的主客观文本分类结果表.

[4] 何跃, 肖敏, 张月. Data.xlsx. 情感倾向改进前后分类结果表.

[5] 何跃, 肖敏, 张月. Data.xlsx. 是否加入结合话题相关性的主客观分类模型时的情感倾向分类结果表.

收稿日期: 2016-10-17
收修改稿日期: 2017-01-25

# Sentiment Analysis of Trending Topics Based on Relevance

He Yue　Xiao Min　Zhang Yue
(Business School, Sichuan University, Chengdu 610064, China)

**Abstract:** [**Objective**] This paper tries to effectively analyze the sentiment of trending topics with machine learning techniques. [**Methods**] First, we proposed a new classification model based on trending topic relevance to extract subjective microblog posts. Second, we analyzed sentiment tendency with an improved machine learning method. [**Results**] We found that the modified model improved the subjective-objective classification of trending topics. The F-measures were increased by 7.4% and 2.2% respectively. [**Limitations**] More research is needed to study the distribution of data, the particle of emotion and the changes of sentiment trends. [**Conclusions**] Adding topic relevance factor to the model could improve the performance of sentiment analysis of micro-blog posts, and extract tendency of key objects from the trending topics, which provides intelligence for micro-blog marketing.

**Keywords:** Trending Topic　Subjective-Objective Classification　Emotion Orientation Classification　TF-IDF-SIM　Machine Learning